Modèle Linéaire Généralisé Hiérarchique Gamma-Poisson à 3 facteurs aléatoires - Application au contrôle de qualité

Florence Loingeville, Julien Jacques, Cristian Preda, Philippe Guarini, Olivier Molinier

Modèle Linéaire Généralisé Hiérarchique Gamma-Poisson à 3 facteurs aléatoires - Application au contrôle de qualité

Florence Loingeville [1,2,4] , Julien Jacques [1,3] , Cristian Preda [1,2] , Philippe Guarini [4] & Olivier Molinier [4]

[1] *Inria Lille - Nord Europe - florence.loingeville@inria.fr*
[2] *Laboratoire Paul Painlevé / Université de Lille 1 - cristian.preda@polytech-lille.fr*
[3] *Université Lumière Lyon 2 - julien.jacques@univ-lyon2.fr*
[4] *AGLAE Hallennes-lez-Haubourdin - philippe.guarini@association-aglae.fr, olivier.molinier@association-aglae.fr*

Résumé. Le dénombrement de particules dans une phase homogène est idéalement représenté par la loi de Poisson. En pratique, il s'avère pourtant que la dispersion des résultats de dénombrements de germes est supérieure à celle attendue d'après le modèle de Poisson. Nous proposons dans ce travail un Modèle Linéaire Généralisé Hiérarchique Gamma-Poisson à trois facteurs aléatoires permettant d'estimer les dispersions induites par les différents facteurs d'un essai interlaboratoires.

Mots-clés. HGLM, Gamma-Poisson, surdispersion, h-vraisemblance

Abstract. Ideally, bacterial counts in an homogeneous phase is represented by the Poisson distribution. However, practicly, dispersion of results of counts have proved to be higher than the dispersion expected from the Poisson model. We propose here a Poisson-Gamma Hierarchical Generalized Linear Model, with three random factors, in order to estimate the amounts of dispersion induced by the different factors of a proficiency test.

Keywords. HGLM, Poisson-Gamma, overdispersion, h-likelihood

1 Introduction

Les laboratoires qui effectuent des analyses de l'environnement ou dans le domaine de la biologie médicale doivent procéder à des contrôles de qualité interne et externe, afin d'assurer le suivi de leur procédure analytique. La comparaison interlaboratoires, exercice qui consiste à soumettre un même essai à plusieurs établissements, est l'outil utilisé pour la mise en œuvre du contrôle externe de qualité. L'organisation optimale des essais interlaboratoires peut être résumée par le plan d'expérience suivant : un matériau à analyser est envoyé à chaque laboratoire participant à l'essai, sous la forme de deux échantillons A et B. Chaque échantillon est séparé en deux réplications (A1, A2 et B1, B2) par le laboratoire, qui réalise ensuite dans des conditions de répétabilité les mesures demandées sur les

quatre réplications dont il dispose. En microbiologie, la mesure sera un dénombrement de particules (germes). À partir des mesures reçues de la part de l'ensemble des laboratoires, trois critères de qualité sont évalués:

- La capacité d'un laboratoire à répéter ses analyses (écarts entre A1 et A2, B1 et B2).
- L'hétérogénéité des préparations envoyées aux laboratoires (écarts entre A et B).
- La justesse des mesures des laboratoires.

L'outil statistique de base pour évaluer ces trois critères de contrôle qualité est l'analyse de variance, qui consiste à décomposer la variance totale des résultats de mesures en une variance inter-laboratoires et une variance intra-laboratoire, laquelle pouvant elle même être décomposée en une variance inter-échantillons (écart entre A et B) et intra-échantillon (écart entre A1 et A2 puis entre B1 et B2). L'objectif est alors d'évaluer la significativité des différences entre les mesures sur la base de cette décomposition.

Dans un premier travail (Loingeville, 2014), nous avons proposé une méthode de test de significativité des effets Laboratoire et Flacon. Nous proposons dans ce papier d'estimer la dispersion des résultats de mesure inter-laboratoires, inter-flacons et intra-flacon, et de classer les laboratoires participant à un essai. En section 2, nous transcrivons le problème de surdispersion en microbiogie en terme statistique. Nous modélisons ensuite le problème en section 3, et proposons une méthode d'estimation en section 4. Enfin, des résultats expérimentaux illustrent la pertinence de la méthode proposée.

2 Expression de la surdispersion en microbiologie

La loi de Poisson décrit la dispersion idéale des résultats de dénombrements en microbiologie. Pourtant, en pratique, lorsqu'une erreur de mesure s'ajoute à l'incertitude liée au dénombrement, l'égalité entre espérance et variance caractéristique de la loi de Poisson n'est plus vérifiée. La loi de Poisson surdispersée est alors représentée par la loi binomiale négative (Bliss, CI, Fisher, RA, 1953).

La paramétrisation classique de la loi Binomiale Négative $BN(n, p)$ (Saporta, 2006) étant peu intelligible en microbiologie, on opère un premier changement de variable:

$$n = \frac{\lambda}{K-1} \qquad\qquad p = \frac{1}{K}$$

où λ est un réel positif correspondant au niveau de charge bactérienne, et K est un réel strictement supérieur à 1 correspondant à la sur-dispersion.

Lorsque l'on calcule les incertitudes de mesure, il est d'usage d'additionner l'effet des composantes. En chimie, la loi normale s'applique et on additionne ainsi les variances. En microbiologie, l'utilisation du ratio variance/moyenne est alors peu explicite. On peut donc encore améliorer la formulation de la surdispersion, en introduisant un paramètre de surdispersion, u^2, issu d'un coefficient de variation, CV^2 (Norme ISO 13843):

$$CV^2 = \frac{V(X)}{E(X)^2} = \frac{1}{\lambda} + u^2 \qquad\qquad u^2 = \frac{K-1}{\lambda} = \frac{1}{n}$$

Nous pouvons ainsi obtenir un modèle additif en coefficients de variation.

La distribution binomiale négative correspond à un mélange Gamma-Poisson. En effet, soit Y une variable aléatoire distribuée suivant une loi de Poisson de paramètre Z, où Z est elle-même une variable aléatoire de distribution Gamma $Z \sim \Gamma(n, \theta)$. Y suit alors une distribution Binomiale Négative $BN(n, p)$ où $p = 1/(\theta + 1)$ (Plackett, 1981). En utilisant la paramétrisation NB 2 (Cameron et Trivedi, 1998), et le paramètre u^2 introduit ci-dessus, $Z \sim \Gamma(\frac{1}{u^2}, u^2)$. Cette paramétrisation est alors particulièrement explicite en microbiologie, puisque $E[Z] = 1$, et $V[Z] = u^2$.

3 Modélisation du problème

Les Modèles Linéaires Généralisés (GLM) étendent les modèles linéaires en terme de loi de probabilité, puisque la classe des distributions est généralisée à toute la famille exponentielle, et en terme de lien à la linéarité, par l'utilisation d'une fonction de lien (Nelder, Wedderburn, 1972). Ces modèles ont donné naissance aux Modèles Linéaires Généralisés Mixtes (GLMM), dans lesquels le prédicteur linéaire peut comporter, outre les effets fixes, une ou plusieurs composantes aléatoires gaussiennes. Puis Lee et Nelder (1996) ont proposé une classe de Modèles Linéaires Généralisés Hiérarchiques (HGLM). Il s'agit de modèles dans lesquels les composantes aléatoires proviennent de distributions arbitraires, et notamment de la distribution conjuguée à celle de la variable observée.

Dans le contexte des essais interlaboratoires, on note y_{ijk} le résultat du dénombrement effectué sur la réplication $k \in \{1, 2\}$ du flacon $j \in \{1, 2\}$ par le laboratoire $i \in \{1...a\}$. Considérons le modèle à lien logarithmique suivant:

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} \quad \begin{cases} i = 1, ..., a \\ j = 1, 2 \\ k = 1, 2 \end{cases}$$

Ceci est équivalent à un modèle où la distribution de y_{ijk} sachant les composantes aléatoires α_i,β_{ij} et γ_{ijk}, suit une loi de Poisson de paramètre λ_{ijk}:

$$E(y_{ijk}|\alpha_i, \beta_{ij}, \gamma_{ijk}) = \lambda_{ijk} = e^{\mu}.\, e^{\alpha_i}.\, e^{\beta_{ij}}.\, e^{\gamma_{ijk}} \tag{1}$$

où e^{μ} est un paramètre fixe correspondant à la moyenne globale des dénombrements de germes, et e^{α_i}, $e^{\beta_{ij}}$ et $e^{\gamma_{ijk}}$ sont trois variables aléatoires indépendantes de lois Gamma:

- $L_i = e^{\alpha_i} \sim \Gamma(\frac{1}{u_1^2}, u_1^2)$ représente l'effet du laboratoire i
- $F_{ij} = e^{\beta_{ij}} \sim \Gamma(\frac{1}{u_2^2}, u_2^2)$ représente l'effet du flacon j du laboratoire i
- $R_{ijk} = e^{\gamma_{ijk}} \sim \Gamma(\frac{1}{u_3^2}, u_3^2)$ représente l'erreur de mesure entre réplications d'un flacon

On constate alors que:

$$E[L_i] = E[F_{ij}] = E[R_{ijk}] = 1$$

et que les paramètres d'échelle des variables aléatoires L_i, F_{ij} et R_{ijk} correspondent respectivement aux parts de surdispersion induites par le laboratoire (u_1^2), le flacon (u_2^2), et la réplication (u_3^2):

$$Var[Li] = u_1^2 \qquad Var[F_{ij}] = u_2^2 \qquad Var[R_{ijk}] = u_3^2$$

Le paramètre λ_{ijk} correspond à un produit de trois variables aléatoires indépendantes de lois Gamma, dont nous ne connaissons donc pas la distribution exacte.
Le modèle (1) correspond à un HGLM Gamma-Poisson à trois facteurs aléatoires.

4 Estimation des paramètres du modèle

L'objectif de ce travail est d'estimer les effets laboratoires L_i, ainsi que les paramètres μ, u_1^2, u_2^2 et u_3^2 du modèle présenté en section 3. Pour estimer les paramètres d'un HGLM, Lee et Nelder (1996) proposent une généralisation de la vraisemblance jointe d'Henderson, la vraisemblance hiérarchique, dite h-vraisemblance. Cette vraisemblance permet d'éviter l'intégration nécessaire lorsque la vraisemblance marginale est utilisée.
Pour le HGLM Gamma-Poisson à trois facteurs aléatoires présenté en section 3, la h-vraisemblance s'écrit:

$$h = L\left(\mu, u_1^2, u_2^2, u_3^2\,;\, y|(L,F,R)\right) \cdot L(L) \cdot L(F) \cdot L(R) \tag{2}$$

où $L(L)$,$L(F)$ et $L(R)$ correspondent aux vraisemblances de variables aléatoires indépendantes de lois Gamma, et $L\left(\mu, u_1^2, u_2^2, u_3^2; y|(L,F,R)\right)$ est la vraisemblance de $y|(L,F,R)$.
A partir de cette h-vraisemblance, Pawitan (2001) propose une méthode d'estimation de μ, des effets laboratoire α_i, ainsi que des coefficients de surdispersion u_1^2, u_2^2, et u_3^2 à l'aide d'un algorithme backfitting.

5 Expérimentation et résultats

Dans cette section, nous présentons les résultats de la méthode proposée sur données simulées, puis sur des données réelles issues de la microbiologie.

5.1 Expérimentation sur données simulées

Pour évaluer la qualité de la méthode proposée, nous la testons dans un premier temps sur données simulées. Pour ce faire, nous utilisons le package "HGLM" de R, dans lequel est implémenté l'algorithme d'estimation dans les HGLM. Nous faisons appel à la fonction *hglm2* afin d'estimer les paramètres d'un HGLM à trois facteurs aléatoires (modèle présenté en section 3).
Nous générons $N = 10000$ jeux de données d'un essai auquel participent $a = 100$ laboratoires. Nous fixons le niveau de concentration moyen $e^{\mu} = e^5$, et choisissons les valeurs

suivantes pour les coefficients de surdispersion: $u_1^2 = 0,20$, $u_2^2 = 0,10$, et $u_3^2 = 0,05$.
Pour chacun des N jeux de données ainsi générés, nous estimons le paramètre fixe du modèle, μ, ainsi que les coefficients de surdispersion, u_1^2, u_2^2 et u_3^2.
Nous calculons ensuite les moyennes sur les N valeurs de μ, u_1^2, u_2^2 et u_3^2 estimées. Puis nous déterminons des intervalles de confiance pour les valeurs estimées de μ, u_1^2, u_2^2 et u_3^2 sur de tels essais. Nous obtenons les résultats suivants:

paramètre	valeur simulée	moyenne des N estimations	intervalle de confiance de l'estimation au risque de 5%
μ	5	4.940425	$[4.83721, 5.04364]$
u_1^2	0.20	0.2078085	$[0.130719; 0.2848979]$
u_2^2	0.10	0.1026843	$[0.06787906; 0.1374896]$
u_3^2	0.05	0.05049129	$[0.03882354; 0.06215904]$

Table 1: Résultats sur N jeux de données simulées

Les résultats des estimations effectuées sur les jeux de données simulés illustrent la pertinence de la méthode proposées pour évaluer les paramètres μ, u_1^2, u_2^2, et u_3^2.

5.2 Application à deux essais interlaboratoires en microbiologie

Nous avons appliqué la méthode proposée à deux jeux de données "métier":

1. Un jeu de données de dénombrement des *Pseudomonas aeruginosa* dans les eaux:
   - Une méthode relativement bien décrite d'un point de vue normatif
   - Un type de colonies assez bien retrouvé par les participants

   Ce jeu de données correspond à un type de dénombrement de difficulté intermédiaire en termes d'exploitation des résultats d'essais interlaboratoires.

2. Un jeu de données de dénombrement de staphylocoques pathogènes dans les eaux:
   - Des disparités interlaboratoires notoires liées à l'utilisation de méthodes différentes (milieu de culture différents) par le panel de participants
   - Une dispersion interlaboratoires relativement large, qui suscite des réflexions de la normalisation AFNOR, en vue d'une tentative d'harmonisation des pratiques

   Il s'agit d'un type de dénombrement plutôt délicat à exploiter.

Nous estimons les paramètres sur ces deux jeux de données à l'aide du package HGLM (Table 2). Les résultats des estimation obtenues, et notamment celles de u_1^2, corroborent notre connaissance métier des jeux de données. Nous obtenons aussi des estimations des effets des laboratoires. Les laboratoires qui surestiment ont un effet supérieure à 1, tandis que ceux qui sous-estiment ont un effet inférieur à 1 (Figure 1).

Jeu de données	$\hat{\mu}$	$\hat{u}_1^2$	$\hat{u}_2^2$	$\hat{u}_3^2$
1- *Pseudomonas aeruginosa*	4.00294	0.0269029	0.0008897	0.0004882
2- staphylocoques pathogènes	3.11840	0.181982	0.008177	0.001525

Table 2: Résultats de l'estimation sur les 2 jeux de données métiers

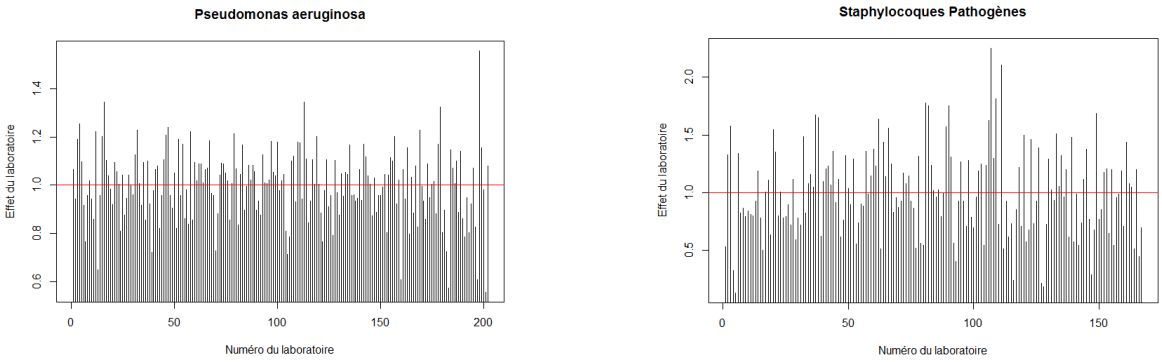

Figure 1: Effets des Laboratoires participants aux essais étudiés

6 Conclusion

Le plan d'expérience des essais interlaboratoires peut être modélisé par un HGLM Gamma-Poisson à trois facteurs aléatoires. Grâce à un algorithme spécifique utilisant la vraisemblance hiérarchique, nous pouvons estimer les paramètres de ce modèle. La méthode proposée permet également de classer les laboratoires prenant part à un essai. Les résultats d'estimation sur données simulées et réelles en confirment la pertinence.

Bibliographie

[1] Bliss, CI, Fisher, RA (1953), *Fitting the Negative Binomial Distribution to Biological Data*, Biometrics, pp.176-200.
[2] A.C. Cameron, P.K. Trivedi (1998), *Regression Analysis of Count Data, New York: Cambridge University Press.*
[3] A.C. Cameron, P. K. Trivedi (1999), *Essentials of Count Data Regression.*
[4] Y. Lee, J.A. Nelder (1996), *Hierarchical Generalized Linear Models*, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 58, No. 4(1996), pp. 619-678.
[5] F. Loingeville, J. Jacques, C. Preda, P. Guarini, O. Molinier (2014), *Analyse de variance à 2 facteurs imbriqués sur données de comptage*, 46ème Journées de Statistique.
[6] ISO 13843 (2000), *Qualité de l'eau - Lignes directrices pour la validation des méthodes microbiologiques.*
[7] J.A. Nelder, J. Wedderburn (1972), *Generalized Linear Models*, Journal of the Royal Statistical Society, Series A (General), Vol. 135, No. 3, pp. 370-384.
[8] Y. Pawitan (2001), All Likelihood Statistical Modelling and Inference Using Likelihood.
[9] R.L. Plackett (1981), *The Analysis of Categorial Data. Griffin*, London.
[10] G. Saporta (2006), *Probabilités, analyse des données et statistique.*